Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online): 2229 5666 Vol. 11, Issue 1 (2018): 21-28

REGRESSION TYPE DOUBLE SAMPLING ESTIMATOR OF POPULATION MEAN USING AUXILIARY INFORMATION

Peeyush Misra
Department of Statistics, D.A.V.(P.G.) College, Dehradun, India
E Mail: dr.pmisra.dav@gmail.com

Received June 14, 2017
Modified January 06, 2018
Accepted April 30, 2018

Abstract

In this paper, a regression type double sampling estimator is introduced to estimate the population mean by using auxiliary information. The expressions for bias and mean squared error are found for the new introduced regression type double sampling estimator of population mean. A comparative study with some of the well-known estimators of the population mean has been done. A separate numerical study is also included to illustrate the performance of the new introduced estimator.

Key Words: Auxiliary Information, Bias, Mean Squared Error, Percent Relative Efficiency.

1. Introduction

We often make use of the information available on an auxiliary variable with the variable under study for improving the efficiency of an estimator. For better understanding one may see Cochran (1977), Des Raj (1968), Mukhopadhyay (2012), and Sukhatme et, al. (1984). According to a double sampling technique, a preliminary large first phase sample of size $n'$is taken from a population of size N and then a second phase sample of size n is drawn from the first phase sample of size n' using simple random sampling without replacement at both the phases. Only the auxiliary variable X is observed at first phase sample of size n' whereas the study variable Y and the auxiliary variable X both are observed at the second phase sample of size n.

Let us denote the population mean of study variable y by $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ andthe population mean of auxiliary variable x by $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. Let

$$\sigma_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2, \qquad \sigma_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$$ and

$\rho = \dfrac{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{\sigma_Y \sigma_X}$ be the population correlation coefficient between y and x.

Also let $\mu_{rs} = \dfrac{1}{N}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^r\left(X_i - \overline{X}\right)^s$, $C_Y^2 = \dfrac{\sigma_Y^2}{\overline{Y}^2}$, $C_X^2 = \dfrac{\sigma_X^2}{\overline{X}^2} = \dfrac{\mu_{02}}{\overline{X}^2}$,

$\rho = \dfrac{\mu_{11}}{\sigma_Y \sigma_X}$, $\beta_2 = \dfrac{\mu_{04}}{\mu_{02}^2}$, $\beta_1 = \dfrac{\mu_{03}^2}{\mu_{02}^3}$, $\gamma_1 = \sqrt{\beta_1}$.

Let the first phase sample of size n' be $\left(x_1', x_2', \ldots, x_n'\right)$ on x and the second phase sample of size n be $\left\{\left(y_1, x_1\right), \left(y_2, x_2\right), \ldots, \left(y_n, x_n\right)\right\}$ on variables $\left(y, x\right)$ with the first phase sample mean $\overline{x}' = \dfrac{1}{n'}\sum_{i=1}^{n'} x'_i$ estimator of population mean $\overline{X}$ and the second phase sample mean $\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$ and mean $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ respectively on y and x.

Let us assume that N is large enough as compared to n so that thef p c terms may be ignored.

To estimate the population mean, anew regression typedouble sampling estimatoris introduced and is given by

$$\hat{\overline{y}} = \overline{y} + b\left(\overline{x}' - \overline{x}\right) + k\left(\frac{s_x^2}{\overline{x}^2} - C_X^2\right) \tag{1.1}$$

where b is an estimate of the change in y when x is increased by unity.

2. Bias and Mean Squared Error (MSE) of the Proposed Estimator

In order to derive the expressions for bias and mean squared error (mse) of the new introduced estimator, let us denote by

$$\begin{aligned} \overline{y} &= \overline{Y}\left(1 + e_0\right) \\ \overline{x} &= \overline{X}\left(1 + e_1\right) \\ \overline{x}' &= \overline{X}\left(1 + e_1'\right) \\ s_{yx} &= e_2 + S_{YX} \\ s_x^2 &= e_3 + S_X^2 \end{aligned} \tag{2.1}$$

so that ignoring finite population correction for simplicity we have

$$E\left(e_0\right) = E\left(e_1\right) = E\left(e_1'\right) = E\left(e_2\right) = E\left(e_3\right) = 0 \tag{2.2}$$

$$E\left(e_0^2\right)=\frac{\mu_{20}}{n\overline{Y}^2}=\frac{1}{n}C_Y^2$$

$$E\left(e_1^2\right)=\frac{\mu_{02}}{n\overline{X}^2}=\frac{1}{n}C_X^2$$

$$E\left(e_1'^2\right)=\frac{\mu_{02}}{n'\overline{X}^2}=\frac{1}{n'}C_X^2$$

$$E\left(e_3^2\right)=\left(\frac{\beta_2(x)-1}{n}\right)S_X^4=\frac{\mu_{02}^2}{n}\left(\frac{\mu_{04}}{\mu_{02}^2}-1\right)$$

$$E\left(e_0e_1\right)=\frac{\mu_{11}}{n\overline{YX}}=\frac{1}{n}\rho C_Y C_X$$

$$E\left(e_0e_1'\right)=\frac{\mu_{11}}{n'\overline{YX}}=\frac{1}{n'}\rho C_Y C_X$$

$$E\left(e_0e_3\right)=\frac{\mu_{12}}{n\overline{Y}}$$

$$E\left(e_1e_1'\right)=\frac{\mu_{02}}{n'\overline{X}^2}=\frac{1}{n'}C_X^2$$

$$E\left(e_1e_2\right)=\frac{\mu_{12}}{n\overline{X}}$$

$$E\left(e_1e_3\right)=\frac{\mu_{03}}{n\overline{X}}$$

$$E\left(e_1'e_2\right)=\frac{\mu_{12}}{n'\overline{X}}$$

$$E\left(e_1'e_3\right)=\frac{\mu_{03}}{n'\overline{X}} \qquad (2.3)$$

The new introduced regression type double sampling estimator represented by $\hat{\bar{y}}$ for estimating the population mean given in (1.1) is

$$\hat{\bar{y}}=\bar{y}+b\left(\bar{x}'-\bar{x}\right)+k\left(\frac{s_x^2}{\bar{x}^2}-C_X^2\right) \qquad (2.4)$$

In terms of e_i's, $i=0,1,2,3$; up to terms of order $O(1/n)$ the above regression type double sampling estimator reduces to

$$\hat{\bar{y}} - \bar{Y} = \bar{Y}e_0 - \beta\bar{X}e_1 + \beta\bar{X}e_1' + \frac{k}{\bar{X}^2}e_3 - \frac{2k}{\bar{X}^2}S_X^2 e_1 + \frac{k}{\bar{X}^2}\left(3S_X^2 e_1^2 - 2e_1e_3\right)$$
$$+ \frac{\bar{X}}{S_X^2}\beta\left(e_1e_3 - e_1'e_3\right) - \frac{\bar{X}}{S_X^2}\left(e_1e_2 - e_1'e_2\right) \qquad (2.5)$$

where $\beta = \dfrac{S_{YX}}{S_X^2}$.

The expression for bias of $\hat{\bar{y}}$ up to terms of order $O(1/n)$ is given by

$$\text{Bias}\left(\hat{\bar{y}}\right) = \left\{E\left(\hat{\bar{y}}\right) - \bar{Y}\right\} = \frac{k}{n\bar{X}^4}\left(3S_X^2\mu_{02} - 2\bar{X}\mu_{03}\right) + \left(\frac{1}{n} - \frac{1}{n'}\right)\frac{1}{S_X^2}\left(\beta\mu_{03} - \mu_{12}\right) \qquad (2.6)$$

The expression for mean squared error of $\hat{\bar{y}}$ up to terms of order $O(1/n)$ is given byMSE $\left(\hat{\bar{y}}\right) = \left\{E\left(\hat{\bar{y}}\right) - \bar{Y}\right\}^2$

$$= \bar{Y}^2E\left(e_0^2\right) + \beta^2\bar{X}^2E\left(e_1^2\right) + \beta^2\bar{X}^2E\left(e_1'^2\right) - 2\beta^2\bar{X}^2E\left(e_1e_1'\right)$$
$$- 2\beta\overline{Y}\overline{X}E\left(e_0e_1\right) + 2\beta\overline{Y}\overline{X}E\left(e_0e_1'\right) + \frac{k^2}{\bar{X}^4}E\left(e_3^2\right) + \frac{4k^2}{\bar{X}^4}S_x^4E\left(e_1^2\right)$$
$$- \frac{4k^2}{\bar{X}^4}S_X^2E\left(e_1e_3\right) + \frac{2k\bar{Y}}{\bar{X}^2}E\left(e_0e_3\right) - \frac{4k\bar{Y}}{\bar{X}^2}S_X^2E\left(e_0e_1\right)$$
$$- \frac{2\beta k}{\bar{X}}E\left(e_1e_3\right) + \frac{4\beta kS_X^2}{\bar{X}}E\left(e_1^2\right) + \frac{2\beta k}{\bar{X}}E\left(e_1'e_3\right) - \frac{4\beta kS_X^2}{\bar{X}}E\left(e_1e_1'\right)$$

using values of the expectation given in (2.2) and (2.3), the above expression reduces to

$$\text{MSE}\left(\hat{\bar{y}}\right) = \frac{\mu_{20}}{n} + \beta^2\mu_{02}\left(\frac{1}{n} - \frac{1}{n'}\right) - 2\beta\mu_{11}\left(\frac{1}{n} - \frac{1}{n'}\right) + \frac{\mu_{02}^2}{n\bar{X}^4}\left(\frac{\mu_{04}}{\mu_{02}^2} - 1\right)k^2$$
$$+ \frac{4S_X^4\mu_{02}}{n\bar{X}^6}k^2 - \frac{4S_X^2\mu_{03}}{n\bar{X}^5}k^2 + \frac{2\mu_{12}}{n\bar{X}^2}k - \frac{4S_X^2\mu_{11}}{n\bar{X}^3}k$$
$$- \frac{2\beta\mu_{03}}{\bar{X}^2}\left(\frac{1}{n} - \frac{1}{n'}\right)k + \frac{4\beta S_X^2\mu_{02}}{\bar{X}^3}\left(\frac{1}{n} - \frac{1}{n'}\right)k \qquad (2.7)$$

which attains the minimum for the optimum value

$$k=\frac{\left\{\frac{4S_X^2\mu_{11}}{n\overline{X}^3}+\frac{2\beta\mu_{03}}{\overline{X}^2}\left(\frac{1}{n}-\frac{1}{n'}\right)-\frac{2\mu_{12}}{n\overline{X}^2}-\frac{4\beta S_X^2\mu_{02}}{\overline{X}^3}\left(\frac{1}{n}-\frac{1}{n'}\right)\right\}}{2\left\{\frac{\mu_{02}^2}{n\overline{X}^4}\left(\frac{\mu_{04}}{\mu_{02}^2}-1\right)+\frac{4S_X^4\mu_{02}}{n\overline{X}^6}-\frac{4S_X^2\mu_{03}}{n\overline{X}^5}\right\}} \tag{2.8}$$

Substituting the above value of k given by (2.8) in (2.7),the expression for minimum mean squared error of $\hat{\bar{\bar{y}}}$ is given by

$$\text{MSE}\left(\hat{\bar{\bar{y}}}\right)_{\min}=\frac{\mu_{20}}{n}+\left(\frac{1}{n}-\frac{1}{n'}\right)\beta^2\mu_{02}-2\left(\frac{1}{n}-\frac{1}{n'}\right)\beta\mu_{11}$$

$$-\frac{\left[2C_X^2\left\{\frac{\overline{X}\mu_{11}}{n}-\beta\overline{X}\mu_{02}\left(\frac{1}{n}-\frac{1}{n'}\right)\right\}+\left\{\beta\mu_{03}\left(\frac{1}{n}-\frac{1}{n'}\right)-\frac{\mu_{12}}{n}\right\}\right]^2}{\frac{1}{n}\left\{\left(\mu_{04}-\mu_{02}^2\right)+4C_X^2\left(\overline{X}^2C_X^2\mu_{02}-\overline{X}\mu_{03}\right)\right\}} \tag{2.9}$$

3. Efficiency Comparison

(i) The general estimator of mean in case of SRSWOR is given by $\hat{\bar{\bar{y}}}_{wor}=\bar{y}$ with $MSE\left(\hat{\bar{\bar{y}}}_{wor}\right)=\frac{\mu_{20}}{n}$ (3.1)

(ii) The usual double sampling regression estimator is given by $\bar{y}_{ld}=\bar{y}+b\left(\bar{x}'-\bar{x}\right)$ with $MSE\left(\bar{y}_{ld}\right)=\frac{\mu_{20}}{n}-\left(\frac{1}{n}-\frac{1}{n'}\right)\frac{\mu_{11}^2}{\mu_{02}}$

(3.2)

Hence it is clear from (2.9), (3.1)and (3.2) that the new introduced estimator $\hat{\bar{\bar{y}}}$ is more efficient than theestimators $\hat{\bar{\bar{y}}}_{wor}$ and $\bar{y}_{ld}$ in the sense of having lesser MSE.

4. Empirical Study

To illustrate the performanceof $\hat{\bar{\bar{y}}}$, let us consider the following data.

Population I: Mukhopadhyay (2012, Page Number - 104)

y : Quality of raw materials (in lakhs of bales)

x : Number of labourers (in thousands)

$\mu_{02}=9704.4475, \mu_{20}=90.95,\ \mu_{11}=612.725,\ \mu_{12}=93756.3475,\ \mu_{03}=988621.5173,$

μ_{40}= 35456.4125, μ_{21}= 11087.635, μ_{22}= 2893630.349, μ_{30}= 1058.55, μ_{04}= 341222548.2,
$\bar{y}$ = 41.5, $\bar{x}$ = 441.95, S_x = 98.51115419, S_y = 9.536770942, ρ =0.652197067,
$\beta_2(y)$= 4.286367314, $\beta_2(x)$= 3.623231573, C_X = 0.22290113, C_Y =0.229801709,
β =0.063138576, n = 20, n' =35 (say).

$MSE\left(\hat{\bar{y}}_{wor}\right)$= 4.5475, $MSE\left(\bar{y}_{ld}\right)$= 3.718501766 and $MSE\left(\hat{\bar{y}}\right)\min$ = 2.973655581.

PercentRelative Efficiency of the estimator $\hat{\bar{y}}$ over $\hat{\bar{y}}_{wor}$ = 152.9262511.

PercentRelative Efficiencyof the estimator $\hat{\bar{y}}$ over $\bar{y}_{ld}$ = 125.048166.

Population II: Murthy (1967, Page Number - 398)
y : Number of absentees
x : Number of workers

μ_{02} = 1299.318551, μ_{20} = 42.13412655, μ_{11} = 154.6041103, μ_{12} = 5086.694392,
μ_{03} = 32025.12931, μ_{40} = 11608.18508, μ_{21} = 1328.325745, μ_{22} = 148328.4069,
μ_{30} = 425.9735118, μ_{04} = 4409987.245, $\bar{y}$ = 9.651162791, $\bar{x}$ = 79.46511628,
S_x = 36.04606151, S_y = 6.491080538, ρ = 0.660763765, $\beta_2(y)$= 6.53877409,
$\beta_2(x)$= 2.612197776, C_X = 0.453608617, C_X = 0.672569791, β = 0.118988612,
n = 43, n' =50 (say).

$MSE\left(\hat{\bar{y}}_{wor}\right)$= 0.979863408, $MSE\left(\bar{y}_{ld}\right)$= 0.919969037 and
$MSE\left(\hat{\bar{y}}\right)\min$ = 0.919483397.

PercentRelative Efficiency of the estimator $\hat{\bar{y}}$ over $\hat{\bar{y}}_{wor}$ = 106.5667321.

PercentRelative Efficiency of the estimator $\hat{\bar{y}}$ over $\bar{y}_{ld}$ = 100.0528166.

Population III: Singh and Chaudhary (1997, Page Number - 176)
y : Total number of guava trees
x : Area under guava orchard (in acres)

μ_{02} = 12.50056686, μ_{20} = 187123.9172, μ_{11} = 1377.39858, μ_{12} = 4835.465464,
μ_{03} = 37.09863123, μ_{40} = 1.48935E+11, μ_{21} = 712662.4414, μ_{22} = 8747904.451,
μ_{30} = 100476814.5, μ_{04} = 540.1635491, $\bar{y}$ = 746.9230769, $\bar{x}$ = 5.661538462,
S_x = 3.535614072, S_y = 432.5782209, ρ = 0.900596235, $\beta_2(y)$= 4.253426603,
$\beta_2(x)$= 3.456733187, C_X = 0.624497051, C_Y = 0.579146949, β = 110.1868895,
n = 13, n' =30 (say).

$MSE\left(\hat{\bar{y}}_{wor}\right)=$ 14394.14747, $MSE\left(\bar{y}_{ld}\right)=$ 7778.476942 and $MSE\left(\hat{\bar{y}}\right)\min=$ 7774.98443

PercentRelative Efficiency of the estimator $\hat{\bar{y}}$ over $\hat{\bar{y}}_{wor}$ = 185.1341003.

PercentRelative Efficiency of the estimator $\hat{\bar{y}}$ over $\bar{y}_{ld}$ = 100.0449199.

Population IV: Singh and Chaudhary (1997, Page Number: 154-155)
y : Number of milch animals in survey
x : Number of milch animals in census

μ_{02} = 431.5847751, μ_{20} = 270.9134948, μ_{11} = 247.3944637, μ_{12} = 3119.839406, μ_{03} = 5789.778954, μ_{40} = 154027.4827, μ_{21} = 2422.297374, μ_{22} = 210594.3138, μ_{30} = 2273.46265, μ_{04} = 508642.4447, $\bar{y}$ = 1133.294118, $\bar{x}$ = 1140.058824, S_x = 20.77461853, S_y = 16.45945002, ρ = 0.723505104, $\beta_2(y)$ = 2.098635139, $\beta_2(x)$ = 2.730740091, C_X = 0.018222409, C_Y = 0.014523547, β = 0.573223334, n = 17, n' =30 (say).

$MSE\left(\hat{\bar{y}}_{wor}\right)=$ 15.93609, $MSE\left(\bar{y}_{ld}\right)=$ 12.32127 and $MSE\left(\hat{\bar{y}}\right)\min=$ 11.85602717.

PercentRelative Efficiency of the estimator $\hat{\bar{y}}$ over $\hat{\bar{y}}_{wor}$ = 134.4133891.

PercentRelative Efficiency of the estimator $\hat{\bar{y}}$ over $\bar{y}_{ld}$ = 103.9240628.

5. Conclusions

(i) From (2.9) it is clear that the new introduced regression typedouble sampling estimator is more efficient than the estimator $\hat{\bar{y}}_{wor}$ based on simple random sampling when no auxiliary information is used and is also more efficient than the usual double sampling regression estimator $\bar{y}_{ld}$ of mean where the auxiliary informationalready in use.

(ii) From (2.8), the mean square error of the estimator $\hat{\bar{y}}$ is minimized for the optimum value

$$k=\frac{\left\{\frac{4S_X^2\mu_{11}}{n\bar{X}^3}+\frac{2\beta\mu_{03}}{\bar{X}^2}\left(\frac{1}{n}-\frac{1}{n'}\right)-\frac{2\mu_{12}}{n\bar{X}^2}-\frac{4\beta S_X^2\mu_{02}}{\bar{X}^3}\left(\frac{1}{n}-\frac{1}{n'}\right)\right\}}{2\left\{\frac{\mu_{02}^2}{n\bar{X}^4}\left(\frac{\mu_{04}}{\mu_{02}^2}-1\right)+\frac{4S_X^4\mu_{02}}{n\bar{X}^6}-\frac{4S_X^2\mu_{03}}{n\bar{X}^5}\right\}} \tag{5.1}$$

For practical purposes the optimum value involving some unknown parameters may not be known in advance, hence the alternative is to replace the unknown parameters of the optimum value by their unbiased estimators giving estimator depending upon estimated optimum value.

Acknowledgement

The author is very thankful to the Editor in chief Dr. Vinod Kumar and anonymous referees for providing some useful comments on an earlier draft of the paper.

References

1. Cochran, W.G. (1977): Sampling Techniques, 3rd edition, John Wiley and Sons, New York.
2. Des Raj (1968): Sampling Theory, McGraw- Hill, New York.
3. Mukhopadhyay, Parimal (2012): Theory and Methods of Survey and Sampling, 2nd edition, PHI Learning Private Limited, New Delhi, India.
4. Sheela Misra, Singh, R. K. and Shukla, A. K. (2013): Modified regression approach in prediction of finite population mean using known coefficient of Variation, Journal of Reliability and Statistical Studies, Vol. 6, Issue 1, p. 59-67.
5. Subramani, J. and Kumarapandian, G. (2012): A class of modified linear regression estimators for estimation of finite population mean, Journal of Reliability and Statistical Studies, Vol. 5, Issue 2, p. 01- 10.
6. Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. And Asok, C. (1984): Sampling Theory of Surveys with Applications, 3rd Edition, Ames, Iowa (USA) and Indian Society of Agricultural Statistics, New Delhi, India.